Early Prediction of Academic Performance in a Database Course Using Moodle Interaction Logs

Antonio Balderas[1,*,†], Juan Antonio Caballero-Hernández[1,†], Manuel Palomo-Duarte[1,†], Andrés Muñoz[1,†] and Juan Manuel Dodero[1,†]

[1]*Universidad de Cádiz, Escuela Superior de Ingeniería, Avda. República Saharaui, s/n, 11510 Puerto Real, Cádiz, Spain*

Abstract

Academic failure and high dropout rates remain persistent challenges in STEM education, particularly in core disciplines like Computer Science. Using Moodle interaction logs for a compulsory course on Database Systems (183,917 log records from 131 students), we engineered 16 behavioral features that capture engagement, regularity, resource diversity, and temporal activity patterns. Four classification models were trained and evaluated using 5-fold cross-validation and an independent test set: Support Vector Machine, Random Forest, k-Nearest Neighbors, and Logistic Regression. The best-performing models reached an AUC of up to 0.812 (Logistic Regression) and 0.809 (Random Forest) on the independent test set. Notably, Random Forest demonstrated a recall of 81.2% for at-risk students, suggesting its potential suitability as the basis for an Early Warning System (EWS). Feature importance analysis reveals that resource access, total engagement, and formative quiz interaction are the strongest predictors of the final outcome. Furthermore, activity within the first four weeks ranks among the top six predictors, providing empirical evidence that early behavioral signals carry significant predictive power. These results suggest that identification of at-risk students is highly feasible and actionable within the first month of instruction, offering a robust foundation for automated pedagogical interventions.

Keywords

Learning Analytics, Academic Performance Prediction, Engagement Patterns, Moodle, Classification, Early Warning Systems

1. Introduction

Student dropout and academic failure represent persistent challenges in STEM education, with significant consequences for both individuals and institutions [1]. Learning Management Systems (LMS), such as Moodle, generate rich interaction logs that capture fine-grained behavioral traces of student engagement throughout a course. Learning Analytics (LA) leverages these digital footprints to support data-informed decision-making in educational contexts [2]. While LMS activity records have been used to analyze a wide range of educational behaviors, from skill assessment to academic integrity [3, 4], the present work focuses on the predictive modeling of academic performance, a priority research area in current LA [5].

Early Warning Systems (EWS) constitute one of the most practically relevant applications of LA: by detecting at-risk students before critical assessment deadlines, instructors can intervene in a timely and targeted manner. However, the effectiveness of EWS depends critically on the quality of the behavioral features extracted from the raw log data, the robustness of the predictive models used, and the temporal window within which reliable predictions can be made.

Database Systems design and programming are core skills in Computer Science and Software Engineering curricula [6], but they often present significant cognitive hurdles for undergraduate students. In particular, the formulation of SQL queries has been shown to yield low success rates, and students exhibit recurring misconceptions rooted in the declarative nature of the language and the complexity

LASI Spain 26: Learning Analytics Summer Institute Spain 2026, Zamora, May 07–08, 2026

*Corresponding author.
†These authors contributed equally.
antonio.balderas@uca.es (A. Balderas); juanantonio.caballero@uca.es (J. A. Caballero-Hernández); manuel.palomo@uca.es (M. Palomo-Duarte); andrez.munoz@uca.es (A. Muñoz); juanma.dodero@uca.es (J. M. Dodero)
0000-0003-0026-7410 (A. Balderas); 0000-0001-9012-8961 (J. A. Caballero-Hernández); 0000-0001-6002-0319 (M. Palomo-Duarte); 0000-0002-8491-4592 (A. Muñoz); 0000-0002-4105-5679 (J. M. Dodero)

of database schemas [7]. Previous technological interventions in this course, such as a conversational agent to support SQL learning [8] and LA approaches to detect fraudulent behavior in online database systems exams [9], provided positive results in addressing these challenges. The present study continues this line of research by applying predictive modelling to the same course context, with the aim of identifying at-risk students early enough to enable timely pedagogical intervention.

This paper reports on a preliminary study conducted in a Database Systems course at a Spanish university during the academic year 2023/24. Our research questions are as follows.

- **RQ1:** Can Moodle interaction logs reliably predict student academic performance (pass/fail) in a Database Systems course?
- **RQ2:** Which behavioral features are the most predictive of the academic outcome?
- **RQ3:** How early can reliable predictions be made?

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the data set and the feature engineering process. Section 4 presents the experimental setup. Section 5 reports results. Section 6 discusses the findings and limitations. Section 7 concludes and outlines future work.

2. Related Work

The use of LMS log data for academic performance prediction has been extensively studied in the past decade. Romero and Ventura [2] provide a comprehensive survey of Educational Data Mining (EDM) techniques applied to e-learning environments, including classification, clustering, and sequence mining approaches.

Building on these foundations, several studies have demonstrated that engagement metrics derived from Moodle logs, such as login frequency, resource access, and forum participation, are highly predictive of final grades [10]. Importantly, the predictive value of LMS interactions depends on how they are classified and on the course delivery mode: Agudo-Peregrina et al. [11] found that interactions correlate significantly with academic performance in online courses, but not in face-to-face settings supported by a VLE. Building on this, Hernández-García et al. [12] proposed the CILC categorization scheme for LMS interactions, showing that context-specific classifications improve explanatory power and highlighting that the relationship between digital traces and performance is also sensitive to instructional design. The present study addresses this concern by grounding the feature engineering process in the specific learning design of the target course, a face-to-face Database Systems course in which Moodle is used primarily for resource dissemination and formative assessment.

Other studies have focused on the temporal dimension of prediction, questioning how early in the course models can achieve actionable accuracy. Jayaprakash et al. [13] showed that predictive precision tends to stabilize after the first four to six weeks, supporting the feasibility of early intervention. However, many existing studies are limited by their focus on single institutional cohorts without prospective validation, restricting the generalizability of their findings [14].

This study presents a systematic comparison of multiple classifiers and evaluates the predictive weight of features derived from different temporal windows in a university-level Database Systems course.

3. Dataset and Feature Engineering

This section details the data collection process and the subsequent transformation of raw logs into predictive variables.

Table 1
Dataset Summary

Parameter	Value
Total log records	183,917
Academic year	2023/24
Students (after exclusions)	131
Failed students	82 (62.6%)
Passed students	49 (37.4%)
Moodle components	Resource, Quiz, Forum, Task, System
Prediction task	Binary classification (pass/fail)

3.1. Data Sources

The target course is a compulsory Database Systems subject delivered face-to-face at a Spanish university, in which Moodle is used primarily for resource dissemination, formative quizzes, and assignment submission. Students are expected to engage with the platform regularly throughout the semester alongside in-person lectures and laboratory sessions.

The data set consists of Moodle interaction logs from the Database Systems course (2023/24 academic year) at a Spanish university (Table 1). After anonymization, the log file contains 183,917 records that span the entire academic year. Each record includes a timestamp, anonymized user identifier, event context, Moodle component (Resource, Quiz, Forum, Task, System) and event name.

Academic results were obtained from official course records. After excluding non-presenting students ($n = 12$) and three repeating students who bypassed the continuous assessment pathway and attended only the final exam, generating no Moodle activity during the course and thus introducing systematic noise incompatible with log-based modeling, the final dataset comprises 131 students used for the binary classification task (pass vs. fail, threshold: 5.0/10): 82 (62.6%) who failed and 49 (37.4%) who passed. All exclusion criteria were applied prior to any model training.

3.2. Feature Engineering

The raw log records were aggregated at the student level to produce 16 behavioral features grouped into four categories.

- **Engagement volume:** total accesses, distinct active days, and per-component access counts (Resource, Quiz, Forum, Task, System).
- **Diversity of engagement:** number of distinct resources visited.
- **Temporal patterns:** activity span in days, regularity (ratio of active days to total span), proportion of late-night accesses (22:00–07:00), proportion of weekend accesses, and days from course start to first login.
- **Early prediction proxies:** cumulative accesses in the first 4, 8, and 12 weeks of the course.

4. Experimental Setup

The data set was divided into a training set (80%, $n = 105$) and a test set (20%, $n = 26$) using stratified sampling to preserve the proportions of the classes. Model selection was performed using 5-fold stratified cross-validation on the training set. The final performance was assessed on the test set that was left out, which was not used at any point during the development of the model.

Four classifiers were evaluated using the Orange Data Mining platform [15]: Logistic Regression (interpretable baseline), Random Forest, k-Nearest Neighbors (kNN), and Support Vector Machine (SVM), all with default hyperparameters. Area Under the Curve (AUC) was selected as the primary metric given the class imbalance (62.6%/37.4%); Classification Accuracy (CA), F1 score, Precision, and

Recall were reported as secondary metrics. Feature importance was assessed using Information Gain, Gini index, and Random Forest impurity-based scores.

More advanced techniques, such as gradient boosting or deep learning approaches, were not explored in this preliminary study; the selected models serve as transparent, interpretable baselines that facilitate comparison with the existing LA literature and are well-suited to the available sample size.

5. Results

This section presents the findings derived from the application of EDM techniques to Moodle interaction logs. The analysis is divided into the comparative evaluation of classifier performance, a detailed examination of the detection capacity for at-risk students, and the ranking of the importance of the extracted behavioral features.

5.1. Model Comparison

Table 2 reports cross-validation and external test set performance for all four classifiers.

Table 2
Model Performance Metrics: 5-fold Cross-Validation (Training) vs. Independent Test Set

Model	AUC (CV)	AUC (Test)	CA (CV)	CA (Test)
SVM	0.804	0.775	0.771	0.692
Random Forest	0.772	**0.809**	0.686	0.769
Logistic Regression	0.788	0.812	0.752	0.692
kNN	0.738	0.803	0.667	**0.808**

The experimental results (Table 2) demonstrate robust predictive capacity in all classifiers. Random Forest and Logistic Regression emerged as the top performers in the external test set, achieving AUC scores of 0.809 and 0.812, respectively.

A notable finding is the performance increase observed on the test phase compared to cross-validation. Although Random Forest yielded an AUC of 0.772 during training, it scaled to 0.809 on the held data. This trend is even more pronounced in the kNN model, which achieved the highest Classification Accuracy (CA = 0.808) on the test set.

These metrics suggest that the models effectively captured stable behavioral signals that generalized, and even improved, when applied to unseen student profiles. The Logistic Regression performance confirms a strong linear relationship between Moodle engagement and academic success, providing a transparent and reliable baseline. Simultaneously, the high kNN accuracy indicates that at-risk students tend to exhibit highly clustered behavioral patterns, which is a crucial finding for the design of targeted interventions. Given the size of the external test set (n = 26), the consistent performance across both evaluation protocols reinforces the model's reliability for an Early Warning System integration.

5.2. Confusion Matrix Analysis

Table 3 presents the confusion matrix for the Random Forest classifier on the independent test set (n = 26). This model was selected for detailed analysis due to its superior balance between AUC (0.809) and predictive stability.

The analysis shows that the Random Forest classifier achieved a recall of 81.2% for the "Fail" class, successfully identifying 13 out of the 16 students who actually failed. The model recorded a balanced error rate, with 3 false negatives (at-risk students predicted as passing) and 3 false positives (passing students predicted to fail). These results indicate a high sensitivity for the detection of academic risk, maintaining a precision of 81.2% in the "Fail" predictions.

Table 3
Confusion Matrix — Random Forest (Independent Test Set)

	Predicted: Pass	Predicted: Fail
Actual: Pass	7	3
Actual: Fail	3	13

5.3. Feature Importance

Figure 1 presents the top 8 features ranked according to the random forest impurity score.

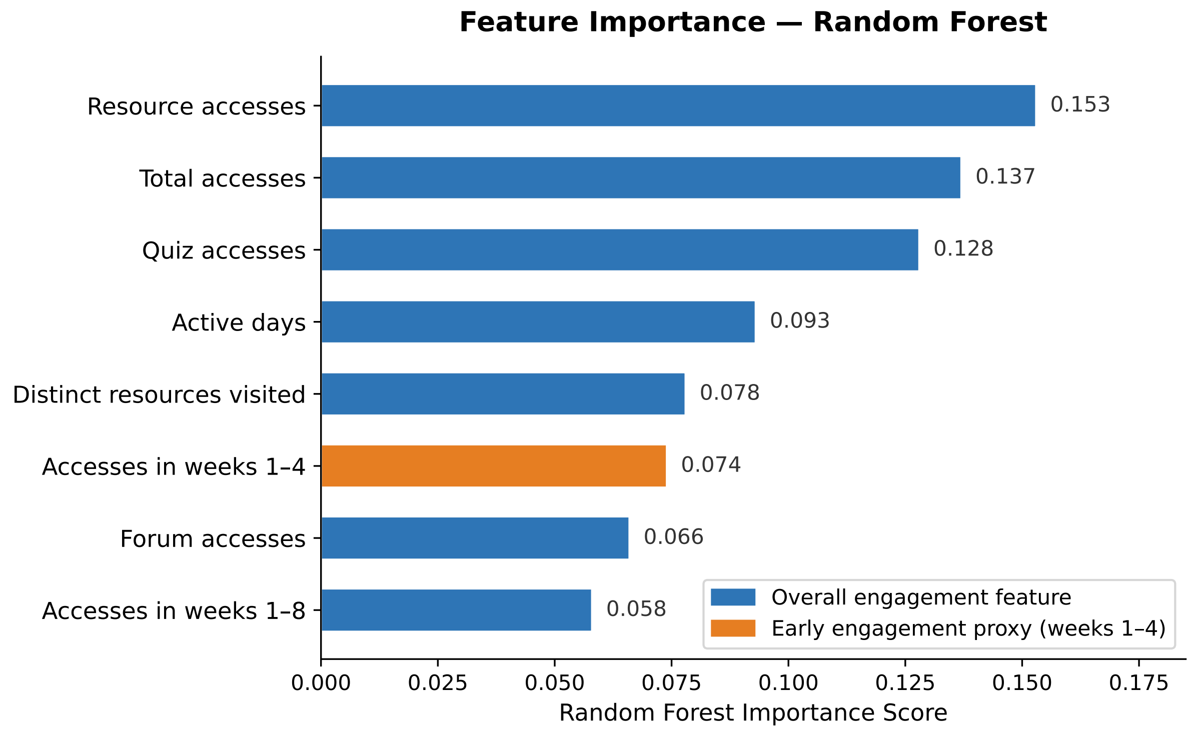

Figure 1: Feature importance scores (Random Forest). The early engagement proxy (accesses in weeks 1–4, shown in orange) ranks 6th overall, supporting the feasibility of early prediction within the first month of instruction.

Resource access emerges as the primary predictor, followed by total engagement and quiz interaction, indicating that the breadth and frequency of content interaction are key determinants of success. Notably, early-stage engagement (accesses in weeks 1–4) ranks 6th overall. This provides empirical evidence that behavioral signals captured within the first month of instruction carry substantial predictive value, directly supporting the feasibility of EWS for timely intervention. In contrast, features related to temporal regularity and specific access timing show lower predictive power, suggesting that the nature of the interaction with Moodle resources is more influential than the specific timing of those actions.

6. Discussion

This section interprets the experimental results in the context of the three research questions (RQs) proposed at the outset of this study.

6.1. Answers to Research Questions

Regarding **RQ1**, Moodle interaction logs reliably predict pass/fail outcomes with an AUC of up to 0.812 (Logistic Regression) and 0.809 (Random Forest), consistent with published benchmarks for comparable datasets in the LA literature [2, 13]. The consistency between the 5-fold cross-validation ($AUC \approx 0.77$) and the independent test set ($AUC \approx 0.81$) demonstrates that the engineered features capture robust behavioral signals. Furthermore, the Random Forest classifier achieved a high recall of 81.2% on the test set, identifying 13 out of 16 at-risk students, which confirms the high sensitivity of the proposed model for academic risk detection.

Regarding **RQ2**, feature importance analysis reveals that resource access (0.153) and overall engagement (0.137) are the main predictors of success. This pattern aligns with self-regulated learning theory [16], where active interaction with materials and formative assessment tools (Quiz accesses, 0.128) serves as a proxy for student commitment. The relatively low predictive power of temporal regularity suggests that the nature and depth of engagement with course content are more critical for academic success than the specific timing or scheduling of study sessions.

Regarding **RQ3**, the inclusion of early-stage activity (accesses in weeks 1–4) among the top 6 most important features (score = 0.074) confirms that early behavioral signals are significant predictors of final results. This finding provides empirical support for the implementation of EWS within the first month of instruction. The high kNN classification accuracy (0.808) on the test set further suggests that at-risk students exhibit early, clustered behavioral patterns that can be identified long before summative assessments occur.

6.2. Limitations

Several limitations must be acknowledged. First, the dataset is restricted to a single Database Systems course at a single institution ($n = 131$). Although high predictive consistency suggests robust behavioral patterns, the findings should be validated across different academic disciplines to ensure broader generalizability. Second, the nature of the course content—focused on Relational Modeling and SQL—implies a high degree of computer-based practice. While Moodle logs effectively capture online interactions, they do not account for offline study behavior or peer-to-peer collaboration, which are common in technical subjects. Third, the class imbalance (62.6%/37.4%) was not corrected in this preliminary analysis; Although the models achieved high sensitivity regardless, future work will explore resampling techniques such as SMOTE [17] to further stabilize the decision boundaries. Fourth, all classifiers were evaluated using default hyperparameters. While this ensures a transparent baseline, systematic hyperparameter tuning and the inclusion of more granular temporal features could yield even higher performance in future iterations of the system.

7. Conclusions and Future Work

This paper demonstrates that Moodle interaction logs can reliably predict academic performance in a face-to-face Database Systems course, with Logistic Regression achieving the highest AUC (0.812) and Random Forest correctly identifying 13 out of 16 at-risk students on an independent test set. The main contributions are:

- A feature engineering framework grounded in the specific learning design of the course, including early-prediction proxies from the first 4, 8 and 12 weeks of activity.
- Empirical evidence that first-4-week engagement already carries significant predictive value, supporting EWS deployment within the first month of instruction.
- Evidence that engagement quality, what students do on Moodle, is a stronger predictor of academic outcome than study schedules or temporal regularity.

This study represents a foundational step in a broader research program. Future work will: (1) conduct a rigorous early prediction analysis restricted to the first 4 and 8 weeks of instruction; (2) implement

SMOTE to address class imbalance; (3) apply systematic hyperparameter optimization; (4) perform prospective validation using data from the 2024/25 and 2025/26 cohorts to ensure temporal stability; and (5) explore clustering approaches to characterize distinct learner behavioral profiles.

Beyond standalone prediction, these indicators enable the integration of risk scores into Multi-Agent Systems (MAS). This study is part of GENIALLE (GENerative Intelligent Agents for digitaL Learning Environments), a framework that leverages LLM-based agents to extend digital platforms with social and proactive capabilities. In this architecture, early risk signals act as triggering events for agents to execute personalized pedagogical interventions. This represents a paradigm shift from passive predictive modeling toward active, autonomous learner support, where AI agents use behavioral data to foster engagement and prevent academic failure in real-time.

Data Availability Statement

The anonymized Moodle interaction logs and the derived feature dataset used in this study are available from the corresponding author upon reasonable request, subject to compliance with applicable data protection regulations.

Acknowledgments

This publication is part of the R&D&i Project PID2023-149674OB-I00, funded by MICIU/AEI/10.13039/501100011033 and ERDF, EU.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist in the design of the data analysis workflow and Gemini (Google) for language editing and structural refinement. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the final content of the publication. The use of these technologies was limited to technical assistance and linguistic improvements; the core conceptual framework, data interpretation, and final conclusions were developed solely by the authors.

References

[1] L. M. H. De Silva, I.-A. Chounta, M. J. Rodríguez-Triana, E. R. Roa, A. Gramberg, A. Valk, Toward an Institutional Analytics Agenda for Addressing Student Dropout in Higher Education: An Academic Stakeholders' Perspective, Journal of Learning Analytics 9 (2022) 179–201. doi:10.18608/jla.2022.7507.

[2] C. Romero, S. Ventura, Educational data mining and learning analytics: An updated survey, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery 10 (2020) e1355. doi:10.1002/widm.1355.

[3] A. Balderas, L. de-la Fuente-Valentin, M. Ortega-Gomez, J. M. Dodero, D. Burgos, Learning Management Systems Activity Records for Students' Assessment of Generic Skills, IEEE Access 6 (2018) 15958–15968. doi:10.1109/ACCESS.2018.2816987.

[4] P. M. Moreno-Marcos, J. Barredo, P. J. Muñoz-Merino, C. D. Kloos, Statoodle: A Learning Analytics Tool to Analyze Moodle Students' Actions and Prevent Cheating, in: Responsive and Sustainable Educational Futures. EC-TEL 2023, volume 14200 of *Lecture Notes in Computer Science*, Springer, Cham, 2023, pp. 736–741. doi:10.1007/978-3-031-42682-7_70.

[5] I. D. Stanciu, A. Hernández-García, M. A. Conde, N. Nistor, Decoding a decade. Trends and evolution in learning analytics: A comprehensive synthesis, Computers in Human Behavior 165 (2025) 108526. doi:10.1016/j.chb.2024.108526.

[6] D. Miedema, E. Aivaloglou, G. Fletcher, Identifying SQL Misconceptions of Novices: Findings from a Think-Aloud Study, in: Proceedings of the 17th ACM Conference on International Computing Education Research (ICER '21), ACM, New York, 2021, pp. 295–305. doi:10.1145/3446871.3469759.
[7] T. Taipalus, M. Siponen, Errors and Complications in SQL Query Formulation, ACM Transactions on Computing Education 18 (2018) 15. doi:10.1145/3231712.
[8] R. Pérez-Mercado, A. Balderas, A. Muñoz, J. F. Cabrera, M. Palomo-Duarte, J. M. Dodero, Chatbot-SQL: Conversational agent to support relational database query language learning, SoftwareX 22 (2023) 101346. doi:10.1016/j.softx.2023.101346.
[9] A. Balderas, M. Palomo-Duarte, J. A. Caballero-Hernández, J. M. Dodero, Learning Analytics to Detect Evidence of Fraudulent Behaviour in Online Examinations, International Journal of Interactive Multimedia and Artificial Intelligence 7 (2021) 241–249. doi:10.9781/ijimai.2021.10.007.
[10] M. Mbodila, F. Elegbeleye, Evaluating student activity and learning on Moodle: A data-driven analysis and insights of usage reports, Studies in Learning and Teaching 6 (2025) 556–569. doi:10.46627/silet.v6i2.623.
[11] Ángel F. Agudo-Peregrina, S. Iglesias-Pradas, M. Ángel Conde-González, Ángel Hernández-García, Can we predict success from log data in VLEs? Classification of interactions for learning analytics and their relation with performance in VLE-supported F2F and online learning, Computers in Human Behavior 31 (2014) 542–550. doi:https://doi.org/10.1016/j.chb.2013.05.031.
[12] Ángel Hernández-García, C. Cuenca-Enrique, L. Del-Río-Carazo, S. Iglesias-Pradas, Exploring the relationship between LMS interactions and academic performance: A Learning Cycle approach, Computers in Human Behavior 155 (2024) 108183. doi:https://doi.org/10.1016/j.chb.2024.108183.
[13] S. M. Jayaprakash, E. W. Moody, E. J. M. Lauria, J. R. Regan, J. D. Baron, Early alert of academically at-risk students: An open source analytics initiative, Journal of Learning Analytics 1 (2014) 6–47. doi:10.18608/jla.2014.11.3.
[14] G. Akçapınar, D. Taşalp, The impact of learning analytics dashboard intervention on student interaction and academic achievement in online learning environments, Interactive Learning Environments (2026) 1–16. doi:10.1080/10494820.2026.2628248.
[15] J. Demšar, T. Curk, A. Erjavec, v. Gorup, et al., Orange: Data mining toolbox in Python, Journal of Machine Learning Research 14 (2013) 2349–2353. URL: https://dl.acm.org/doi/10.5555/2567709.2567736.
[16] P. R. Pintrich, The role of goal orientation in self-regulated learning, in: M. Boekaerts, P. R. Pintrich, M. Zeidner (Eds.), Handbook of Self-Regulation, Academic Press, San Diego, 2000, pp. 451–502. doi:10.1016/B978-012109890-2/50043-3.
[17] A. Anggrawan, H. Hairani, C. Satria, Improving SVM Classification Performance on Unbalanced Student Graduation Time Data Using SMOTE, International Journal of Information and Education Technology 13 (2023) 289–295. doi:10.18178/ijiet.2023.13.2.1806.